Diffusion de systèmes de préférences par confrontation de points de vue, vers une simulation de la Sérendipité

Guillaume Surroca[1], Philippe Lemoisson[1,2],
Clément Jonquet[1], Stefano A. Cerri[1]

[1] Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM),
Université de Montpellier & CNRS
nom@lirmm.fr

[2] UMR Territoires, Environnement, Télédétection et Information Spatiale (TETIS),
CIRAD, Montpellier, France
philippe.lemoisson@cirad.fr

Résumé : Le Web d'aujourd'hui est formé, entre autres, de deux types de contenus que sont les données structurées et liées du Web sémantique et les contributions d'utilisateurs du Web social. Notre ambition est d'offrir un modèle pour représenter ces contenus et en tirer communément avantage pour l'apprentissage collectif et la découverte de connaissances. En particulier, nous souhaitons capturer le phénomène de Sérendipité (i.e., de l'apprentissage fortuit) à l'aide d'un formalisme de représentation des connaissances subjectives où un ensemble de points de vue forment un graphe de connaissances interprétable de façon personnalisée. Nous établissons une preuve de concept sur la capacité d'apprentissage collectif que permet ce formalisme appelé Viewpoints en construisant une simulation de la diffusion de connaissances telle qu'elle peut exister sur le Web grâce à la coexistence des données liées et des contributions des utilisateurs. A l'aide d'un modèle comportemental paramétré pour représenter diverses stratégies de navigation Web, nous cherchons à optimiser la diffusion de systèmes de préférences. Nos résultats nous permettent d'identifier les stratégies les plus adéquates pour l'apprentissage fortuit et d'approcher la notion de Sérendipité. Une implémentation du noyau du formalisme Viewpoints est disponible ; le modèle sous-jacent permet l'indexation de tous types de jeux de données.

Mots-clés : représentation des connaissances, découverte et diffusion de connaissances, Sérendipité, ingénierie des connaissances centrée utilisateurs, apprentissage, intelligence collective, Web 2.0, agents.

1 Introduction

Depuis que le Web 2.0 a démocratisé la création, le partage et la recommandation de contenus notamment grâce aux réseaux sociaux, aux blogs et aux forums et depuis que les technologies du Web sémantique ont commencé à structurer la connaissance du Web, deux formes de contenus s'y sont dégagées. Celles-ci sont différentes dans leurs façons de s'établir et dans leurs niveaux de structuration. D'une part les plateformes contributives du Web social permettent la production d'une profusion de données peu ou pas structurées mais au cycle d'évolution et d'entretien très rapide (e.g., les folksonomies [15]). D'autre part des connaissances très structurées sont constituées par consensus par des cercles d'experts, acteurs de la construction du Web sémantique (e.g., les ontologies [8] ou les données liées [3]). Avec l'approche Viewpoints, notre ambition est de créer un formalisme de représentation des connaissance qui intègre aussi bien les jeux de données structurées et liées du Web sémantique que l'abondance d'interactions du Web social afin de tirer le double avantage de la structuration qui caractérise les jeux de données du Web sémantique et de la vitesse d'évolution et d'entretien des connaissances partagées sur le Web social comme l'envisageait Gruber dans [7] ou Ankolekar dans [2]. Notre objectif est de remettre au centre du modèle de

représentation des connaissances la contribution des agents du Web (humains ou artificiels) sous forme de points de vue reliant des ressources (identifiées par une URI). Nous nous posons les questions suivantes :

1. Quelles sont les stratégies de navigation sur le Web qui permettent la diffusion optimale des systèmes de préférences des utilisateurs ?
2. Comment positionner les conditions propres à l'apprentissage fortuit, c'est-à-dire à la Sérendipité, dans l'étude des systèmes de préférences ?

Nous parlerons de système de préférences d'un agent pour identifier l'ensemble des goûts ou attirances qu'il exprime sous forme de relations de proximité ou de distance entre ressources du Web. Dans une précédente contribution [10] nous avons démontré la capacité d'apprentissage d'une base de connaissances construite à partir d'une première ébauche de notre formalisme. Toutefois, cette preuve de concept ne se basait que sur une modélisation très pauvre du comportement des agents qui naviguaient au hasard au sein de la base de connaissances pour y contribuer ; nous nous intéressions alors seulement à la satisfaction des utilisateurs sans prendre en compte leurs systèmes de préférences. Dans une autre contribution, nous avons montré comment Viewpoints permet la recherche et la découverte de connaissances grâce à un prototype de recherche de publications scientifique [19]. Dans la modélisation du comportement des agents que nous proposons aujourd'hui, nous incluons un paramètre d'« ouverture à la Sérendipité » qui est la propension d'un agent à s'orienter vers des ressources hors de son système de préférences pour guider sa recherche ; cela nous permet d'évaluer la diffusion des systèmes de préférences selon si agent est plutôt ouvert d'esprit ou plutôt focalisé sur ce qu'il connait et préfère. A partir de cette modélisation, nous construisons une simulation dans laquelle nous créons des règles de comportement individuel (niveau microscopique) et observons l'effet sur l'apprentissage collectif et la diffusion des systèmes de préférences (niveau macroscopique). Cette simulation donne les grandes lignes de l'effet de l'utilisation de Viewpoints pour encapsuler des données du Web sémantique et social.

L'article est structuré de la façon suivante : la section 2 pose le contexte et les inspirations de notre approche en présentant la notion de Sérendipité dans un cadre informatique. L'état de l'art présente aussi un positionnement de Viewpoints par rapport aux autres approches de représentation des connaissances. Ensuite, nous rappelons brièvement le formalisme Viewpoints dans la Section 3. La section 4 explicite notre modèle du comportement des utilisateurs du Web et notre représentation des systèmes de préférences, montre comment nous simulons l'évolution du Web en tant que graphe de connaissances et expose nos hypothèses concernant l'impact des stratégies individuelles de navigation. La section 5 présente une simulation mettant en œuvre trois agents (les princes de Serendip) contribuant à tour de rôle à un graphe de connaissances 'jouet' construit avec des ressources de différentes formes, tailles et couleurs, puis ouvre sur une discussion des résultats au regard de nos hypothèses et de notre problématique. La section 6 conclue et présente des perspectives possibles à ce travail.

2 Etat de l'art et inspirations

2.1 Représentation des connaissances

Plusieurs travaux se sont positionnés sur le rapprochement du Web sémantique et du Web social [7]. Nous positionnons notre approche par rapport à cet héritage de la manière suivante : il s'agit d'une représentation des connaissances qui en plus d'intégrer l'agent comme présenté dans [13, 15] le considère de manière centrale. Qui plus est, il s'agit d'une représentation des connaissances qui considère le point de vue comme micro-expression des sémantiques individuelles tel que [11]. Cependant, le mécanisme d'arbitrage et de confrontation des points de vue ne fait appel à aucune contribution supplémentaire. Ensuite, l'accent est mis sur l'émergence au sein du graphe biparti, de même que [1, 16] qui étudiaient la possibilité de l'émergence d'une représentation des connaissances collective dans une vision « bottom-up », à partir des interactions d'un système. Pour finir, nous définissons une distance

métrique sur l'ensemble des ressources formé par les fournisseurs, descripteurs et supports de connaissance alors que les distances sémantiques que l'on trouve dans la littérature s'appliquent à des sous-classes homogènes telle que des distances entre tags ou concepts d'ontologie [9].

2.2 La Sérendipité ou l'apprentissage fortuit

Ce mot est dérivé d'un ancien comte persan *Les trois princes de Serendip* [14] Merton écrit à propos du phénomène de Sérendipité qu'il « concerne l'expérience assez générale de l'observation d'une donnée non-anticipée, anormale et stratégique qui devient l'occasion du développement d'une nouvelle théorie, ou l'extension d'une théorie existante. » Plus récemment, Perriault disait « L'effet Serendip (…) consiste à trouver par hasard et avec agilité une chose que l'on ne cherche pas. On est alors conduit à pratiquer l'inférence abductive, à construire un cadre théorique qui englobe grâce à un bricolage approprié les informations jusqu'alors disparates » [17]. Nous notons que la notion de hasard (termes 'hasard' ou 'accident' dans les définitions) est importante dans le phénomène de Sérendipité. Toutefois, elle ne dépend pas uniquement du dans 'divin jet de dés' [6] et n'a lieu qu'à la frontière de ce que l'on sait déjà[1]. Ainsi, les apprentissages fortuits sont grandement facilités lorsque les nouvelles connaissances se situent au voisinage de connaissances existantes et qu'elles peuvent être interprétées par quelqu'un qui connait ce voisinage. Nous partageons cette vision selon laquelle la préparation, l'entraînement et la connaissance ne garantissent pas la découverte par Sérendipité mais elles la rendent plus probable. Nous pouvons ainsi parler de *zone proximale de Sérendipité* de façon similaire à la notion de zone proximale de développement [21]. Nous montrerons par la suite comment nous avons traduit *l'ouverture à la Sérendipité* dans notre modèle.

La Sérendipité existe d'autant plus sur le Web au vu de l'immense quantité de données qu'il contient et des chances que l'on a de s'y perdre. Nous pouvons donc parler d'apprentissage sérendipiteux sur le Web tel qu'expliqué ci-après. La recherche de connaissances par l'apprentissage sérendipiteux peut aboutir par chance ou comme sous-produit d'une tâche principale [4]. Par exemple, un utilisateur fait une recherche initiale qui le mène, au fur et à mesure de l'exploration des résultats, sur une trajectoire tangente non-prévue initialement qui in fine s'avère plus productive que sa première requête. Dans de tels cas Bowles écrit que l'apprentissage sérendipiteux a lieu [4]. C'est exactement le phénomène que nous modélisons et observons dans notre simulation section 4 à l'aide de stratégie de navigation. En addition, selon Allen Tough, presque 80% de l'apprentissage est informel et non planifié [18]; la navigation sérendipiteuse est une « loterie intellectuelle (…) peu de probabilité mais gros gain potentiel » [12]. Dans ce dernier travail, le parallèle avec notre approche Viewpoints est évident : « nous gagnons aussi de nouveaux points de vue ou associations pour notre problème en parcourant des sources alternatives utilisant des outils, des techniques et des structures de données différentes ».

De ces réflexions sur la Sérendipité nous retiendrons les principes suivants : (i) Les esprits préparés sont mieux disposés pour reconnaître la découverte fortuite (principe de Pasteur) c'est-à-dire ce qui est dans la *zone proximale de Sérendipité*. (ii) Le hasard joue un rôle fondateur car il permet de générer assez de chaos pour permettre l'innovation et la découverte. Comme écrit Toubia, «… l'anarchie de la pensée augmente la probabilité d'avoir des idées créatives …» [20].

La Sérendipité intéresse de plus en plus les acteurs des systèmes de recommandation car même si la précision des recommandations est importante leur variété l'est tout autant. La Sérendipité permet d'aller au-delà de ce que faisaient les systèmes de recommandation en créant la surprise, la variété et la nouveauté dans les résultats proposés. D'ailleurs plusieurs

[1] Comme le disait Pasteur « la chance favorise les esprits préparés » et elle a en l'occurrence favorisé celui d'Alexandre Fleming qui s'il n'avait pas été expert n'aurait pas reconnu la pénicilline comme résultat accidentel de son travail qui consistait à l'origine à faire des cultures de staphylocoques dans le but d'étudier l'effet antibactérien du lysozyme. Ses boîtes de Pétri furent contaminées accidentellement et il se rendit compte qu'autour des champignons qui avait contaminé ses boîtes les staphylocoques ne poussaient plus.

systèmes de recommandation ont commencé à mettre en œuvre ces principes. Par exemple, StumbleUpon.com permet par exemple à ses utilisateurs de « trébucher » sur une ressource du Web au hasard tout en appliquant le principe de Pasteur car la recommandation se fait en fonction de ses activités récentes ou des goûts qu'il a exprimé ce qui nécessite de modéliser ses préférences. L'utilisateur est donc préparé à « trébucher ». La recommandation basée sur un folksonomie présentée dans [22] permet par exemple aux utilisateurs en associant des livres à des tags en plus de dépasser la classification traditionnelle de rajouter de nouveaux livres dans la *zone proximale de Sérendipité* d'autres utilisateurs. Cependant, à notre connaissance, en dehors de [5] proposant un cadre théorique du phénomène de Sérendipité, la littérature sur la formalisation et la mesure de ce phénomène est pratiquement inexistante. Il n'existe pas aujourd'hui de modèle de la Sérendipité.

3 Le formalisme Viewpoints

Viewpoints est un formalisme de représentation des connaissances subjectives, c'est à dire que toute relation de proximité ou distance entre deux *ressources* est exprimée par un *agent* sous la forme d'un *viewpoint* typé reliant ces deux *ressources*. L'exploitation de ces *viewpoints* est assujettie à une évaluation a posteriori, selon une *perspective* choisie par l'utilisateur/contributeur, en fonction de qui a émis les *viewpoints*, de quand ils ont été émis, de leurs types ou d'autres critères plus complexes. Cela fait de Viewpoints un formalisme de représentation des connaissances centré sur les agents (qui sont eux même des ressources) au sens large : humain ou artificiel (e.g., automate de fouille de données) qui sont tous deux considérés de la même manière. L'ensemble des *ressources* (fournisseurs, descripteurs et supports de connaissances) liées par les points de vue forment le graphe de connaissances. Par exemple, dans [19] nous avions illustré le formalisme dans un prototype de moteur de recherche sur données de publications scientifiques indexées à l'aide des métadonnées bibliographiques (auteurs, articles, mots-clés). Le graphe de connaissances (KG) est un graphe biparti $K_{R, V}$ constitué d'une part d'un ensemble de ressources R et d'un ensemble de *viewpoints* V reliant ces ressources entre elles. Les ressources de R sont soit des agents (fournisseurs de connaissances, c'est à dire créateurs de viewpoints), soit des descripteurs de connaissances (des tags de folksonomies ou bien des concepts d'ontologies) ou bien des supports de connaissances (vidéos, pages Web, message, post, etc.). Un *viewpoint* est un quadruplet ($a \rightarrow \{r_1, r_2\}, \theta, t$) contenant les informations suivantes :

- a, l'*agent* qui a exprimé ce viewpoint
- $\{r_1, r_2\}$, le couple de *ressources* sémantiquement connectées par a
- θ, le type du *viewpoint*, qui va permettre d'interpréter la relation créée
- t, la date de création du *viewpoint*.

Par exemples : `(Guillaume→ {Diffusion de systèmes […] points de vue, acm : Knowledge representation and reasoning}, dc:subject, 27/02/2015)` signifie que l'agent Guillaume associe par la relation DublinCore subject cet article au concept Knowledge representation and reasoning d'ACM. `(Mario→ {Mario, Luigi}, foaf:knows, 1985)` signifie que Mario exprime le fait qu'il connaisse Luigi, il émet donc un viewpoint qui le rapproche de Luigi. Pour identifier le sens des données représentées sous formes de viewpoints, nous réutilisons tant que possible les types existants du Web sémantique.

4 Exploitation des points de vue

L'ensemble des connexions entre deux ressources dues aux différents agents constitue un lien de proximité nommé synapse. La force de cette synapse est fonction de l'agrégation des poids résultant des évaluations de chaque viewpoint. Les deux fonctions d'*évaluation* (map) et d'*agrégation* (reduce) des *viewpoints* sont au cœur de la notion de *perspective* qui permet d'exploiter le graphe de connaissances. C'est-à-dire, que d'un même KG, plusieurs interprétations – des Knowledge Maps (KM) – peuvent être faites qui dépendent de la façon

dont l'agent qui observe évalue et on agrège les viewpoints de chacun. Le KG interprété est un graphe $G_{R, S}$ composé de ressources (R) et de synapses (S). Ainsi, les algorithmes s'exécutant sur KG peuvent être directement adaptés sans effort à des algorithmes de graphe classiques s'exécutant sur KM. La perspective est propre à chaque utilisateur de KG qui décide d'interpréter KG de la manière qu'il souhaite. Par exemple, une perspective simple (telle que présentée dans [19]) donnerait un poids de 1 a tous les viewpoints (map) et calculerait la valeur d'une synapse en faisant une simple somme (reduce). Les deux fonctions d'évaluation et d'agrégation des viewpoints peuvent être étendues à volonté pour correspondre mieux aux usages de notre formalisme.[2] Parmi les algorithmes de graphes nous pouvons mentionner l'algorithme de Dijkstra, du plus court chemin, que nous utilisons dans le calcul de la distance sémantique entre deux ressources quelconques ou l'algorithme de détection de communauté de Louvain que nous utilisons quand nous cherchons une partition de KM. La Figure 1 illustre le processus d'interprétation de KG. Dans la simulation ci-après, nous utilisons : (i) une fonction de voisinage direct qui renvoie pour une ressource r_i toutes les ressources r_j directement reliées par des viewpoints à r_i, ainsi que les poids des synapses liant r_i et r_j ; (ii) une fonction de voisinage indirect qui se base sur l'algorithme de Dijkstra et renvoie pour une ressource r_i toutes les ressources r_j sur tous les chemins partant de r_i et de longueur inférieure ou égale à m (paramètre fixé pour la simulation). Le noyau du formalisme Viewpoints est implémenté en Java et nous utilisons Neo4j pour le stockage de KG. Une interface programmatique (API) existe pour l'indexation de n'importe quel jeu de données[3].

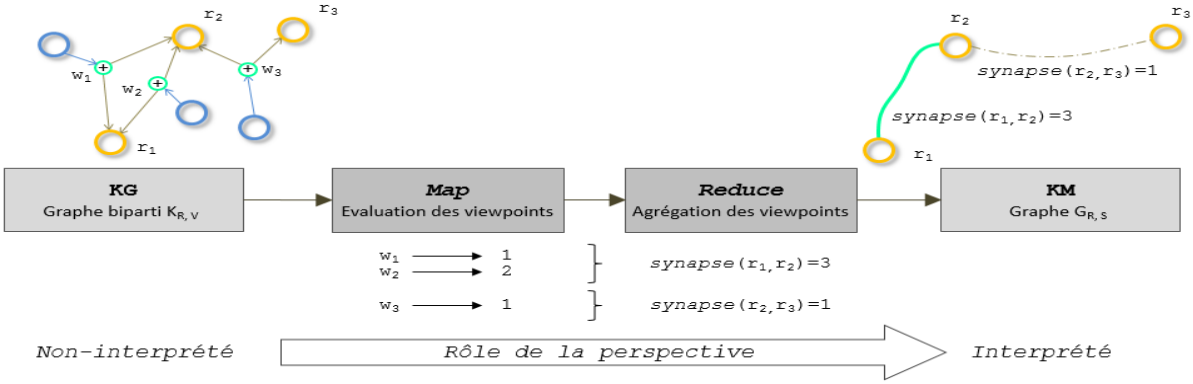

FIGURE 1 – Illustration du processus d'interprétation de KG

5 Simulation des princes de Serendip

Nous souhaitons simuler l'évolution d'une base de connaissances telle que le Web à partir de règles de comportement individuelles qui décrivent les navigations d'agents sur le Web et la diffusion de leurs systèmes de préférences respectifs. Nous commençons par expliquer comment nous représentons les systèmes de préférences dans un graphe de connaissances Viewpoints, puis nous proposons un modèle du comportement simulant différentes stratégies de navigation paramétrables. Ce modèle est fondé sur les calculs de voisinage direct et indirect. Finalement, nous observerons l'effet de cet ensemble de règles individuelles sur le niveau macroscopique de la connaissance représentée dans KG.

5.1 Représentation des systèmes de préférences

Chaque ressource de KG est caractérisée par une forme, une taille et une couleur qui serviront à les rapprocher. Les informations de forme et de taille seront déjà présentes dans le graphe de départ de la simulation ; ces informations simulent les données du Web sémantique. Les informations de couleur des ressources seront ajoutées au fur et à mesure de la simulation

[2] Le fait de concevoir l'exploitation de notre graphe à l'aide d'une approche map/reduce nous ouvre un grand potentiel en termes de traitement de gros volume de données et de passage à l'échelle. Cela n'a pas était formellement mesuré à ce jour, mais nous obtenons d'ores et déjà des performances intéressantes sur les jeux de données que nous avons testés.

[3] https://github.com/sifrproject/viewpoints_kernel

par 3 agents (rouge, vert, bleu), les princes de Serendip, qui connaissent et aiment respectivement une couleur distincte ; ces informations simulent les contributions du Web social. Le système de préférences d'un prince est traduit par les viewpoints qu'il émet pour se rapprocher des ressources de sa couleur ou rapprocher entre elles des ressources de même couleur (la sienne). La diffusion d'un système de préférences est donc équivalente à la diffusion de l'information de couleur dans le graphe. Ainsi, l'apprentissage de la couleur par le graphe représente l'émergence d'une intelligence collective de la communauté. Nous considérons ici deux types de viewpoints : (i) rapprochant deux ressources de même couleur (`vps:knows`) (ii) rapprochant un prince d'une couleur à une ressource de la même couleur (`vps:likes`). Par exemple, si le prince rouge fait une recherche sur une ressource r qui est rouge et obtient parmi les résultats une ressource r' qui est aussi rouge alors il créera les deux types de viewpoints : `(prince rouge→ {prince rouge, r}, vps:likes, τ)` et `(prince rouge→{r, r'}, vps:knows, τ)`. Dans la section suivante nous présenterons les stratégies de navigation dans KG qui permettent à un prince de diffuser la connaissance de sa couleur.

5.2 Modèle comportemental des princes

L'automate à état (Figure 2) décrit le comportement des princes quand ils naviguent dans KG, et diffusent au fur et à mesure de leurs feedbacks (émissions de viewpoints) leurs systèmes de préférences. Plus généralement, cet automate nous permet de décrire le comportement d'un utilisateur explorant le contenu d'une base de connaissances telle que le Web. Nous capturons ainsi des comportements tels que : la requête sur moteur de recherche, l'exploration des résultats, l'exploration des liens inclus dans ces résultats et le retour éventuel au moteur de recherche avec une autre requête, etc. Les probabilités qui conditionnent les transitions dans cet automate dépendent de trois paramètres :

- β, qui est la probabilité de revenir en arrière pendant la navigation, c.-à-d. soit de revenir à la recherche d'origine (état de départ) ou à la dernière recherche effectué (état précèdent).
- μ, qui est le choix parmi les outils de navigation disponibles : soit l'utilisation du moteur de recherche opérant globalement sur le graphe soit l'exploration locale des résultats de proche en proche en suivant les liens qui les connectent.
- σ, qui est la capacité à diriger sa navigation vers des ressources qui n'appartiennent pas forcément à son propre système de préférences : l'ouverture à la Sérendipité.

Dans notre simulation le comportement d'un prince de Serendip correspond à un paramétrage spécifique de β, μ et σ ; nous parlerons de *stratégie de navigation*. Ces stratégies simulent des stratégies de navigation sur le Web (ou autre base de connaissances). La simulation se divise en cycles qui correspondent à des explorations successives de KG. Au début d'un cycle, un prince commence par une interaction avec KG qui simule l'utilisation d'un moteur de recherche : une ressource de KG est sélectionnée aléatoirement et nous utilisons la fonction de voisinage indirect pour obtenir une liste de résultats (autres ressources) triés. A partir des résultats proposés, le prince poursuit (β faible) ou abandonne cette recherche et en fait une nouvelle (β fort). S'il poursuit, il va évaluer (relativement à la couleur correspondant à son système de préférences) ces résultats un par un et opter pour la premier non-visitée en fonction du paramètre σ. Si le prince est ouvert à la Sérendipité (σ fort), alors il ne se dirigera pas systématiquement vers une ressource de même couleur que lui, sinon (σ faible[4]) il privilégiera sa couleur. Ayant choisi une ressource, le prince passera à la prochaine étape de son cheminement, en fonction de μ, soit en faisant une recherche sur cette ressource (μ fort) soit en explorant localement autour de cette ressource (μ faible). La première interaction simule le fait d'ouvrir une page Web après avoir cliqué sur une des URL proposées par le moteur de recherche ; l'interaction suivante simule soit une nouvelle recherche sur par

[4] S'il avait choisi le moteur de recherche Qwant.com il aurait donc commencé par le premier résultat qui lui semble correspondre le plus à ses goûts (σ faible).

exemple le titre de la page, soit le clic sur un lien inclus dans celle-ci. Dans la simulation, un prince dispose d'un budget d'interactions qui diminue à chaque interaction (recherche ou exploration). Ce budget représente la quantité d'effort qu'il est prêt à faire dans sa navigation. Si au moment du retour en arrière il n'y a plus d'étapes précédentes, s'il n'y a plus de ressources non-visitées ou si son budget d'interaction a été dépensé alors le cycle s'achève.

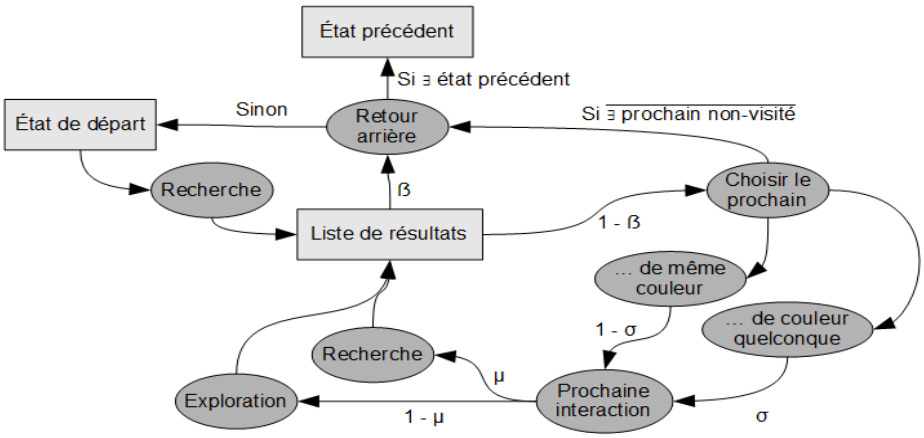

FIGURE 2 – Automate de comportement des princes : stratégies de navigation.

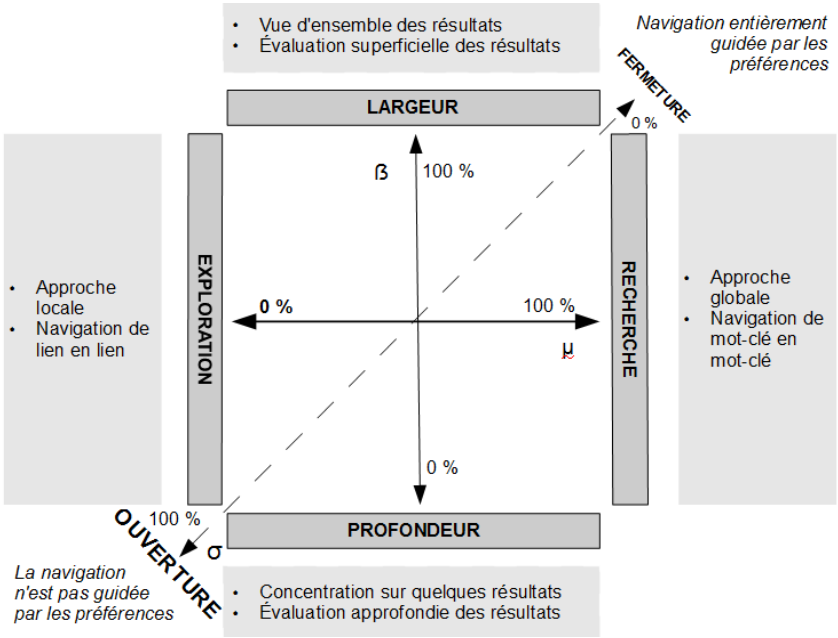

Figure 3 – Différentes stratégies de navigation en fonction des paramètres β, μ et σ.

Nous représentons dans la Figure 3 les trois paramètres relatifs aux stratégies de navigation dans un espace à trois dimensions. Ces stratégies mises en place dans la simulation des princes de Serendip simulent des stratégies de navigation sur le Web. En terme de parcours de graphe plus β est élevé plus on se rapproche d'un parcours en LARGEUR et plus β est faible plus il s'agit d'un parcours en PROFONDEUR. Dans une démarche de recherche d'information le parcours en LARGEUR reviendrait à évaluer de manière superficielle l'ensemble des résultats pour avoir une idée d'ensemble de tous les résultats et l'approche par PROFONDEUR reviendrait plutôt à se concentrer sur ce qui paraîtrait être la meilleure réponse et la creuser plus en profondeur. μ conditionne le style de navigation. Quand μ est élevé on utilise majoritairement des moteurs de RECHERCHE renvoyant des résultats triés et indirectement liés tandis que quand μ est faible on explore de proche en proche en récupérant

des résultats non-triés et directement liés (EXPLORATION). Par exemple, la navigation entre vidéos suggérées sur YouTube est un bon cas de figure d'une exploration de proche en proche tandis que l'utilisation répétitive de Google dans une recherche est plutôt un exemple de parcours en LARGEUR. Nous représentons l'ouverture à la Sérendipité (σ) comme une troisième dimension. Quand σ est grand c'est que l'utilisateur est dans une démarche d'OUVERTURE et qu'il est disposé à cheminer aussi bien parmi des ressources qui correspondent à ses préférences que d'autres ressources qui n'y correspondent pas mais qui pourrait l'amener à la découverte fortuite. Dans le cas opposé (FERMETURE), l'utilisateur parcourt le Web entièrement guidé par ses préférences.

5.3 Déroulement de simulation

5.3.1 Conditions initiales

Un KG de taille déterminée est généré. Les ressources de ce KG sont des ressources caractérisées par une taille (petit, moyen, grand), une forme (carré, cercle, triangle) en plus de posséder une couleur (rouge, vert, bleu). Pour chaque combinaison possible de taille, forme et couleur N ressources sont créées. Il y a donc initialement 27N ressources dans KG. Deux agents artificiels que nous appellerons péons sont ajoutés à KG. L'un d'entre eux partage son appréciation de la notion de forme au graphe de connaissances en reliant tous les couples de ressources de même forme par des viewpoints de type `svp:initial`. L'autre péon fera de même pour la caractéristique de taille. Ainsi, après le passage des péons, KG ne connaît pas la couleur car les ressources ne sont liées que par les deux caractéristiques de taille et de forme. Pour finir la phase d'initialisation trois autres agents sont ajoutés à KG : les princes. Chacun est caractérisé par une couleur unique et a la capacité d'apprécier cette couleur et de partager cette appréciation en émettant de nouveaux viewpoints de type `svp:like` et `svp:knows` dans le graphe de connaissances. Il y a donc une connaissance implicite que les princes sont seuls aptes à partager, par émission de viewpoints de feedback.

5.3.2 Diffusion des systèmes de préférences par confrontation de points de vue

Les paramètres de la simulation sont résumés dans le Tableau 1. Le prince suit le modèle comportemental que nous avons précédemment défini et diffuse ses préférences (la connaissance de sa couleur) en émettant des viewpoints de type svp:like et svp:knows. Le poids associé à chaque type de viewpoint est indiqué dans le Tableau 1. La fonction d'agrégation des viewpoints pour le calcul de la valeur des synapses est la somme. A la fin de chaque cycle les mesures suivantes sont effectuées. Elles nous permettent d'évaluer la diffusion de la connaissance des couleurs dans KG :

- `M1 Couleur X` : Il s'agit du ratio : distance[5] moyenne entre ressources quelconques / distance moyenne entre ressources de couleur X.
- `M2 Couleur X` : Il s'agit de la probabilité d'obtenir au voisinage d'une ressource de couleur spécifique des ressources de la même couleur.

Étant donné le nombre important de paramètres (Tableau 1), nous ne présenterons des résultats obtenus (courbes) que pour certaines simulations, que nous avons jugées les plus significatives pour l'étude des stratégies de navigation. Cependant, nous expliquerons les effets de tel ou tel paramètres dans la section discussion. Dans la suite, les valeurs fixes des paramètres sont précisées dans le Tableau 1.

[5] La mesure de distance employée est une distance aux propriétés métriques (symétrie, séparation et inégalité triangulaire) basée sur le calcul du plus court chemin de Dijsktra (cf. [19]).

TABLE 1 – *Résumé des paramètres de la simulation et de leurs valeurs fixes.*

Catégorie	Paramètre	Valeur (si fixée)	
Paramètres d'échelle	Facteur d'échelle (N)	3	
	Nombre de cycles	100	
	Nombre d'interactions par cycle	50	
Paramètres de perspective	Poids associé aux viewpoints de type `vps:initial`	1	
	… de type `vps:knows`	2	
	… de type `vps:like`	1	
Paramètres de stratégie de navigation	ß		
	µ		
	σ		
Répartition de l'activité	Prince rouge	33%	80%
	Prince vert	33%	10%
	Prince bleu	33%	10%
Paramètres d'algorithme	Borne de distance pour le calcul de voisinage sémantique (m)	2	

5.4 Hypothèses

Au fur et à mesure que les princes contribuent à KG ils partagent leurs appréciations des couleurs avec les autres utilisateurs grâce au mécanisme de feedback. Nous souhaitons observer comment, après leurs contributions, KG aura « appris » au niveau global la notion de couleur qui n'était pas dans les connaissances originalement représentées par les viewpoints. Chaque système de préférences individuel d'un prince devient ainsi, grâce aux viewpoints, une part de la connaissance collective représentée dans KG où il cohabite avec les systèmes de préférences des autres princes. Nous souhaitons expérimenter différentes stratégies de navigation et démontrer que les systèmes de préférences diffusés de façon concurrente ne se neutralisent pas. Nous souhaitons également mesurer l'effet de la Serendipité. Ainsi, nous nous attendons à ce que la mesure M1 augmente, c'est-à-dire à ce que la distance moyenne entre ressources de même couleur décroisse plus vite que la distance moyenne entre ressource quelconques. En effet, les princes rapprochent les ressources de même couleur d'eux-mêmes et les unes des autres, sans jamais rapprocher deux ressources de couleurs différentes. Pour les mêmes raisons, la mesure M2 devrait augmenter aussi car elle reflète la probabilité de trouver une ressource de même couleur dans le m-voisinage d'une ressource.

6 Résultats et discussions

Dans un premier temps, nous faisons varier les stratégies de navigation en conservant la symétrie dans le comportement des trois princes et dans leur répartition de l'activité. Nous observons comment KG « apprend » la couleur rouge en mettant l'accent sur le paramètre σ (ouverture à la Sérendipité). Dans un second temps, nous nous restreignons à deux stratégies de navigation contrastées et jouons sur des répartitions d'activité différentes pour les trois princes ; nous comparons alors les apprentissages respectifs des trois couleurs par KG.

6.1 Impact de l'ouverture à la Sérendipité

Nous commençons par évaluer l'impact de σ sur la diffusion de la couleur rouge grâce aux mesures `M1 Rouge` et `M2 Rouge`. Nous remarquons (Figure 4) que dans le cas d'une utilisation majoritaire du moteur de recherche M1 et M2 croissent plus vite quand l'ouverture

est faible mais qu'inversement quand l'ouverture est élevée elles atteignent des valeurs finales plus élevées. L'ouverture à la Sérendipité permet au final une diffusion plus grande de la connaissance des couleurs. En effet, alors que la recherche renvoie des résultats indirectement liés et permet de créer des viewpoints qui n'avaient pas d'ores et déjà été créés, l'ouverture à la Sérendipité augmente ce potentiel de création de viewpoints nouveaux. Ces nouvelles 'associations' sont des expressions de systèmes de préférences qui n'auraient sans doute pas été créées si la navigation des princes n'avait été guidée que par leurs préférences. Par contraste, nous observons (Figure 5) que dans une approche d'exploration locale ou seuls ne sont renvoyés des résultats directement liés l'ouverture à la Sérendipité n'apporte rien ni en terme de croissance des valeurs M1 et M2, ni en terme de valeur finale obtenue. L'idée, avec une telle stratégie, est d'explorer localement et en profondeur les résultats, ainsi le fait de passer par des résultats moins intéressants en chemin a plutôt tendance à freiner la diffusion des systèmes de préférences. L'effet de μ (outil de navigation) est donc très important sur la Serendipité. Nous nous rendons toutefois compte de la relative homogénéité de notre graphe par rapport à la structure du Web. Nous pensons que la Sérendipité peut apporter en condition réelle un saut qualitatif plus substantiel que celui que nous mesurons sur ce graphe 'jouet'. Dans cette simulation les trois princes sont également actifs (33%) et ß=10%[6]

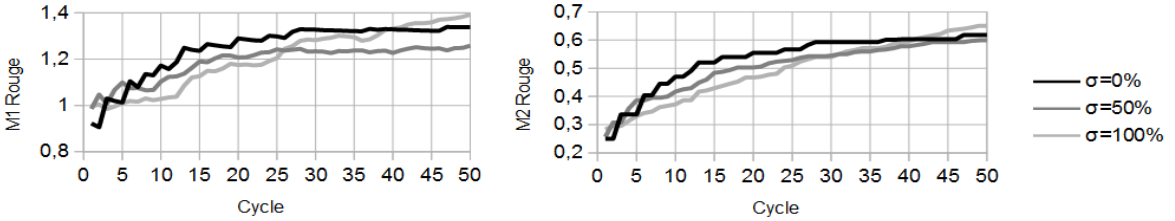

FIGURE 4 – Pour μ=70% et ß=10% (Recherche Profondeur plus ou moins Ouverte).

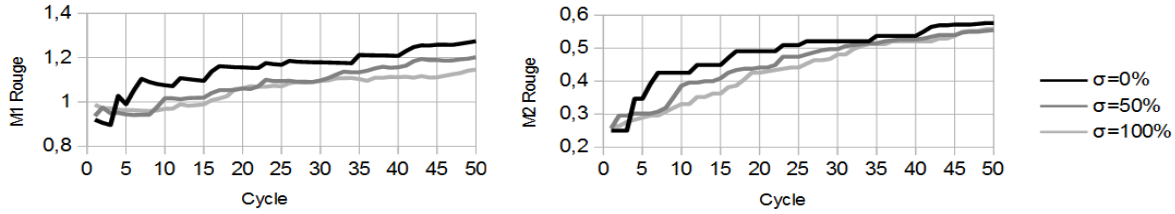

FIGURE 5 – Pour μ=30% et ß=10% (Exploration Profondeur plus ou moins Ouverte).

6.2 Impact de la répartition de l'activité parmi les princes

Dans cette partie nous analysons l'impact de la répartition de l'activité des princes sur la diffusion des couleurs. Pour cela nous observons comparativement `M1 Rouge`, `M1 Vert` et `M1 Bleu` qui évaluent chacune la diffusion d'une couleur dans le graphe. Dans cette simulation, nous faisons varier les probabilités associées aux degrés d'activité respectifs des princes et considérons successivement deux configurations contrastées pour les stratégies de navigation : Recherche Largeur Fermée (μ=80%, ß=40%, σ =10%) et Exploration Profondeur Ouverte (μ=20%, ß=10%, σ =70%). Nous comparons les résultats obtenus pour ces stratégies avec une réparation homogène de l'activité des princes (Figure 6) et avec une répartition non-homogène (Figure 7). Dans les deux cas, nous remarquons que la diffusion de chaque couleur se fait même si la concurrence ralentit cette diffusion. Lorsque les princes sont en concurrence, l'apprentissage d'une couleur se fait bien au détriment d'une autre (lorsque M1 augmente pour un cycle donnée, les autres diminuent) et le prince le plus actif diffuse plus efficacement sa couleur. Cependant, la somme des `M1 Rouge`, `M1 Vert` et `M1 Bleu` finales a une valeur plus élevée quand toutes les connaissances sur les couleurs peuvent être diffusées (Figure 7, la somme des valeurs finales vaut respectivement 3.9 et 3.6) que quand une couleur

[6] Nous avons pu étudier au fur et à mesure de nos simulations que la variation du paramètre ß ne change pas les résultats que nous présentons ci-après. Ainsi, nous le fixons dans toutes les simulations présentées à 10% donnant ainsi priorité aux stratégies en profondeur.

domine dans la diffusion des couleurs (Figure 6, la somme des valeurs finales vaut respectivement 3.6 et 3.5). D'après ces résultats, on peut déduire que les contributions des utilisateurs du Web social ne neutralisent pas celles des autres mais peut les occulter en passant au premier plan.

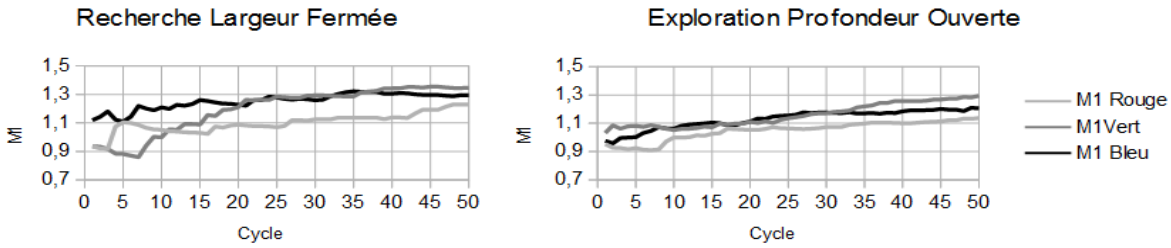

FIGURE 6 – Tous les princes contribuent autant (33%)

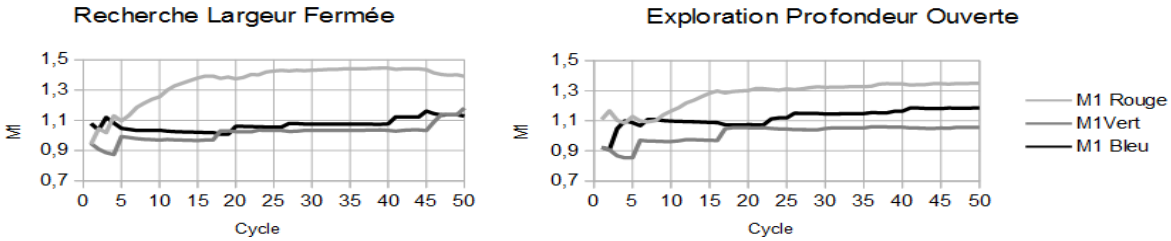

FIGURE 7 – Le prince rouge est plus actif (80%) que les autres (10%).

7 Conclusion et perspectives

Après avoir présenté et positionné notre approche de représentation des connaissances subjectives nous avons étudié le phénomène de Sérendipité et son influence pour le Web d'aujourd'hui. A travers la simulation des princes de Serendip nous présentons un essai de modélisation de la Sérendipité sur le Web. Nous sommes toutefois conscients que ce modèle du comportement des utilisateurs du Web ne rend pas complètement compte de la réalité et de la diversité des méthodes d'explorations du Web. Malgré cela, nous espérons avoir démontré la capacité d'apprentissage du graphe de connaissances de Viewpoints. Les résultats de la simulation nous permettent d'évaluer l'apport de l'ouverture à la Sérendipité dans diverses stratégies de navigation ainsi que son impact sur la diffusion des systèmes de préférences ; nous avons donc consolidé la preuve de concept de Viewpoints en le confrontant à une modélisation d'usage plus réaliste que lors de nos dernières simulations. Toutefois la preuve de concept ultime reste de nous confronter aux vrais usages et données du Web. Nous avons d'ores et déjà commencé la transition vers les cas d'études sur des données réelles du Web social et sémantique en indexant des données cinématographiques contenant 1M de notations de films d'utilisateurs de MovieLens[7]. Nous prévoyons aussi deux cas d'utilisation orientés vers la facilitation de la découverte scientifique transversale dans des contextes de représentation des connaissances agronomiques (Cirad) et biomédicales dans le cadre du projet SIFR (http://www.lirmm.fr/sifr). En outre, en plus d'évaluer l'approche par les usages, nous comptons nous comparer aux algorithmes de recherche d'information en utilisant des benchmarks spécialisés comme on peut en trouver sur LETOR[8] et les mesures de rappel et de précision. Ces cas d'utilisation permettront également de confronter notre approche à de grandes quantités de données et aux problèmes de passage à l'échelle. Pour finir, étant donné que la modélisation des points de vue est le centre de notre approche, nous étudions ce que l'approche Viewpoints pourrait apporter à l'outillage d'e-démocratie censé favoriser l'émergence de projets politiques en facilitant l'accès de chacun à la contribution par l'expression de points de vue.

[7] http://datahub.io/dataset/movielens

[8] http://research.microsoft.com/en-us/um/beijing/projects/letor//

Remerciements

Ce travail a bénéficié des soutiens du Cirad et du projet SIFR financé en partie par le programme JCJC de l'Agence nationale de la Recherche (ANR-12-JS02-01001), l'Université de Montpellier, le CNRS et l'Institut de Biologie Computationnelle de Montpellier.

Références

[1] Aberer, K., Cudr, P., Catarci, T., Hacid, M., Illarramendi, A., Mecella, M., Mena, E., Neuhold, E.J., De, O., Risse, T. and Scannapieco, M. 2004. Emergent Semantics Principles and Issues. *Database Systems for Advanced Applications*. 2, (2004), 25–38.

[2] Ankolekar, A. and Krötzsch, M. 2007. The two cultures: Mashing up Web 2.0 and the Semantic Web. *Proceedings of the 16th international conference on World Wide Web*. 825–834.

[3] Bizer, C., Health, T. and Berners-Lee, T. 2009. Linked Data - The Story So Far. *Semantic Web and Information Systems*. 5, 3 (2009), 1–22.

[4] Bowles, M. 2004. Relearning to E-learn: Strategies for Electronic Learning and Knowledge. *Educational Technology & Society*. 7, 4 (2004), 212–220.

[5] Corneli, J., Pease, A. and Colton, S. 2014. Modelling serendipity in a computational context. *arXiv preprint*. (2014).

[6] Fine, G.A. and Deegan, J.G. 1996. Three principles of Serendip: insight, chance, and discovery in qualitative research. *International Journal of Qualitative Studies in Education*. 9, 4. 434–447

[7] Gruber, T. 2008. Collective knowledge systems: Where the Social Web meets the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*. 6, 1, 4–13.

[8] Karapiperis, S. and Apostolou, D. 2006. Consensus building in collaborative ontology engineering processes. *Journal of Universal Knowledge Management*. (2006), 199–216.

[9] Lee, W.-N., Shah, N., Sundlass, K. and Musen, M. 2008. Comparison of ontology-based semantic-similarity measures. *AMIA, Annual Symposium proceedings 2008*. (Jan. 2008), 384–8.

[10] Lemoisson, P., Surroca, G. and Cerri, S. 2013. Viewpoints : an alternative approach toward Business Intelligence. *e-Challenges e-2013 Conference*. (2013), 8.

[11] Limpens, F. and Gandon, F. 2011. Un cycle de vie complet pour l ' enrichissement sémantique des folksonomies. *Extraction Gestion de Connaissance EGC 2011*. (2011), 1–12.

[12] Marchionini, G. 1997. *Information Seeking in Electronic Environments*.

[13] Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A. and Stumme, G. 2009. Evaluating similarity measures for emergent semantics of social tagging. *Proceedings of the 18th international conference on World wide web - WWW '09*, 641.

[14] Merton, R.K. and Barber, E. 2006. *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science*.

[15] Mika, P. 2007. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*. 5, 1 (Mar. 2007), 5–15.

[16] Noh, T., Park, S., Park, S. and Lee, S. 2010. Learning the emergent knowledge from annotated blog postings. *Web Semantics: Science, Services and Agents on the World Wide Web*..

[17] Perriault, J. 2000. Effet diligence, effet serendip et autres défis pour les sciences de l'information. *Actes du colloque international Pratiques collectives distribuées sur Internet*.

[18] Tough, A. Reflections on the Study of Adult Learning: 1999. .

[19] Surroca, G., Lemoisson, P., Jonquet, C. and Cerri, S.A. 2014. Construction et évolution de connaissances par confrontation de points de vue : prototype pour la recherche d'information scientifique. *IC - 25èmes Journées francophones d'Ingénierie des Connaissances*.

[20] Toubia, O. 2006. Idea Generation, Creativity, and Incentives. *Marketing Science*. 25, 5. 411–425.

[21] Vygotski, L. 1933. Apprentissage et développement: tensions dans la zone proximale. *Paris: La dispute (2ème éd. Augmentée)*. 233, (1933).

[22] Yamaba, H., Tanoue, M. and Takatsuka, K. 2013. On a serendipity-oriented recommender system based on folksonomy. *Artificial Life and Robotics*. 18, 1-2 (2013), 89–94.